ELEVATING FORENSIC INVESTIGATION SYSTEM FOR FILE CLUSTERING

Prashant D. Abhonkar[1], Preeti Sharma[2]

[1]*Department of Computer Engineering, University of Pune SKN Sinhgad Institute of Technology & Sciences, Lonavala, Pune, Maharashtra, India*

[2]*Department of Computer Engineering, University of Pune SKN Sinhgad Institute of Technology & Sciences, Lonavala, Pune, Maharashtra, India*

Abstract

In computer forensic investigation, thousands of files are usually surveyed. Much of the data in those files consists of formless manuscript, whose investigation by computer examiners is very tough to accomplish. Clustering is the unverified organization of designs that is data items, remarks, or feature vectors into groups (clusters). To find a noble clarification for this automated method of analysis are of great interest. In particular, algorithms such as K-means, K-medoids, Single Link, Complete Link and Average Link can simplify the detection of new and valuable information from the documents under investigation. This paper is going to present an tactic that applies text clustering algorithms to forensic examination of computers seized in police investigations using multithreading technique for data clustering.

Keywords- *Clustering, forensic computing, text mining, multithreading.*

----------***----------

1. INTRODUCTION

A very huge increase in crime relating to Internet and computers has caused a growing need for computer forensics. In document clustering computer forensics identifies evidence when computers are used in the police investigations of crimes. In this particular application domain, it usually involves examining the thousands of files per computer. This activity exceeds the expert's ability of analysis and understanding of data.[1.4.5]

In general, for computer forensic analysis we need computer forensic tools that can exist in the form of computer software. Such tools have been developed to help computer forensic investigators in a computer investigation. However, because storage media is growing in size, day by day investigators may have difficulty in locating their points of interest from a large pool of data.[1.5] In addition, the format in which the data is presented may result in disinforming and difficulty for the investigators. As a result, the process of analyzing large volumes of data may consume a very large amount of time. It may happen that data generated by computer forensic tools may be meaningless at times, due to the amount of data that can be stored on a storage medium and the fact that current computer forensic tools are not able to present a visual overview of all the objects (e.g. files) found on the storage medium.[1]

Basically this is paper for the police investigations through forensic data analysis. Clustering algorithms are typically used for examining data analysis, where there is little or no prior knowledge about the data. This is exactly the case in several applications of Computer Forensics, including the one mention in this paper.[1]

Following table gives the various algorithms and their parameters. [1]

Table 1: Summary of algorithms and their parameters [1]

Acronym	Algorithm	Attributes	Distance	Initialization	K-estimate
Kms	*K-means*	Cont. (all)	Cosine	Random	Simp. Sil.
Kms100	*K-means*	100 > TV	Cosine	Random	Simp. Sil.
Kms100*	*K-means*	100 > TV	Cosine	[18]	Simp. Sil.
KmsT100*	*K-means*	100 > TV	Cosine	[18]	Silhouette
KmsS	*K-means*	Cont. (all)	Cosine	Random	Rec. Sil.
Kms100S	*K-means*	100 > TV	Cosine	Random	Rec. Sil.
Kmd100	*K-medoids*	100 > TV	Cosine	Random	Silhouette
Kmd100*	*K-medoids*	100 > TV	Cosine	[18]	Silhouette
KmdLev	*K-medoids*	Name	Lev.	Random	Silhouette
KmdLevS	*K-medoids*	Name	Lev.	Random	Rec. Sil.
AL100	*AverageLink*	100 > TV	Cosine	-	Silhouette
CL100	*CompleteLink*	100 > TV	Cosine	-	Silhouette
SL100	*SingleLink*	100 > TV	Cosine	-	Silhouette
NC	CSPA	Name, Cont. (all)	CSPA	Random	Simp. Sil.
NC100	CSPA	Name,100 > TV	CSPA	Random	Simp. Sil.
E100	CSPA	Cont.100 random	CSPA	Random	Simp. Sil.

100 > TV: 100 attributes (words) that have the greatest variance over the documents
Cont. 100 random: 100 randomly chosen atributes from document content
Cont. (all): all features from document content
Lev.: Levenshtein distance
Simp. Sil.: Simplified Silhpuette
Rec. Sil.: "Recursive" Silhouette
*: Initialization on distant objects
Name: file name

As shown in table there are various algorithm with their parameters like distance which has cosine as well as levenshtein distance which is nothing but a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. The application for levenshtein distance is to in approximate string matching, the objective is to find matches for short strings in many longer texts, in situations where a small number of differences is to be expected. Table also gives the initialization of each algorithm.[1]

The remainder of this paper is organized as follows. Section II presents related work. Section III Architectural diagram of proposed system Section IV concludes the paper. Section V gives references.

2. LITERATURE SURVEY-

The use of clustering has been reported by only few studies in the computer forensics field.[1] Basically, The use of classic algorithm for clustering data is described by most of the studies such as Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, K-means, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM). These algorithms have well-known properties and are widely used in practice.[4]

An integrated environment for mining e-mails for forensic analysis, using classification and clustering algorithms, was presented in [4]. In a related application domain, e-mails are grouped by using lexical, syntactic, structural, and domain-specific features [6]. Three clustering algorithms (K-means, Bisecting K-means and EM) were used. The problem of clustering e-mails for forensic analysis was also addressed, where a Kernel-based variant of K-means was applied [7]. The obtained results were analyzed subjectively, and the result was concluded that they are interesting and useful from an investigation perspective. More recently, a FCM-based method for mining association rules from forensic data was described [3].

In this paper when we talk about computer forensics there are so many tools, algorithms and methods to do it. so this paper presents those algorithms and methods are going to discuss one by one.

Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection uses various algorithms and preprocessing technique for giving result as cluster data. Finally in their conclusion they have shown that, the approach presented by them applies document clustering methods to forensic analysis of computers seized in police investigations. Also, they are reported and discussed with several practical results that can be very useful for researchers and practitioners of forensic computing. More specifically, in their experiments the hierarchical algorithms known as Average Link and Complete Link presented the best results. Despite their usually high computational costs, they have shown that those algorithm are particularly suitable for the studied application domain because the dendrograms that they provide offer summarized views of the documents being inspected, thus being helpful tools for forensic examiners that analyze textual documents from seized computers.[1]

A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering describe one of the central problems in text mining and information retrieval area is text clustering. Performance of clustering algorithms will considerably reject for the high dimensionality of feature space and the inherent data sparsity, two techniques are used to deal with this problem: feature extraction and feature selection. Feature selection methods have been successfully applied to text categorization but seldom applied to text clustering due to the unavailability of class label information. Four unsupervised feature selection methods such as DF, TC, TVQ, and a new proposed method TV were introduced in that paper. Experiments were taken to show that feature selection methods can improves efficiency as well as accuracy of text clustering.[2]

Fuzzy Methods for Forensic Data Analysis is again describes a methodology and an automatic procedure for inferring accurate and easily understandable expert-system-like rules from forensic data. In most data analysis environments the methodology and the algorithms used were proven to be easily implementable. By discussing the applicability of different fuzzy methods to improve the effectiveness and the quality of the data analysis phase for crime investigation the fuzzy set theory would get implemented.[3]

In mining write prints from anonymous e-mails for forensic investigation, basically they are collecting e-mails written by multiple anonymous authors and focusing on the problem of mining the writing styles of those e-mails. The general idea is to first cluster the anonymous e-mail by the Stylometric (Stylometry is the application of the study of linguistic style, usually to written language, but it has successfully been applied to music and to fine-art paintings as well) features and then extract the write print, i.e., the unique writing style, from each cluster.[4]

They have mainly focus on lexical and syntactic features of an e-mail as when we talk about lexical features they are used to learn about the preferred use of isolated characters and words of an individual. Following table gives Some of the commonly used character-based features, these include frequency of individual alphabets (26 letters of English), total number of upper case letters, capital letters used in the beginning of sentences, average number of characters per word, and average number of characters per sentence. To indicates the preference

of an individual for certain special characters or symbols or the preferred choice of selecting certain units the use of such features come in picture. For example most of the people prefer to use '$' symbol instead of word 'dollar', '%' for 'percent', and '#' instead of writing the word 'number'.[4]

Now when we talked about syntactic features, they are also called as style markers which Consist of all purpose function words such as 'though', 'where', 'your', punctuation such as '!' and ':', parts-of-speech tags and hyphenation etc. as shown in table. [4]

Table 2 Lexical And Syntactic Features.

LEXICAL AND SYNTACTIC FEATURES	
Features type	**Features**
Lexical: character-based	1. Character count (N) 2. Ratio of digits to N 3. Ratio of letters to N 4. Ratio of uppercase letters to N 5. Ratio of spaces to N 6. Ratio of tabs to N 7. Occurrences of alphabets (A-Z) (26 features) 8. Occurrences of special characters: < > % j { } [ ]/ \ @ # w þ _ * $ ^ & O (21 features)
Lexical: word-based	9. Token count(T) 10. Average sentence length in terms of characters 11. Average token length 12. Ratio of characters in words to N 13. Ratio of short words (1e3 characters) to T 14. Ratio of word length frequency distribution to T (20 features) 15. Ratio of types to T 16. Vocabulary richness (Yule's K measure) 17. Hapax legomena 18. Hapax dislegomena
Syntactic features	19. Occurrences of punctuations, . ? ! : ; ' " (8 features) 20. Occurrences of function words (303 features

Exploring Data Generated by Computer Forensic Tools with Self- Organizing Maps is again one paper which gives several tools for computer forensic analysis. The demonstration on how unsupervised learning neural network model, the self-organizing map (SOM), can aid computer forensic investigators in decision making and assist them in conducting the analysis process more efficiently during a computer investigation is the main focus of that paper.[5]

This paper talked about the market conditions of tools of computer forensic there are numerous computer forensic tools available on the market. For instance EnCase , Forensic Toolkit and Pro Discover are some of the tools that are available. By considering the difference of these tools some may offer a whole range of functionalities on the other hand some tools are designed with only a single purpose in mind. Examples of these functionalities are advanced searching capabilities, hashing verification, report generation and many more. Some computer forensic tools do provide similar functionalities, but with a different graphical user interface.[5]

The main motto of this paper is to focus on self-organising map (SOM) – which is an approach that allows computer investigators to view (visualise) all the files on the storage medium and assists them in locating their points of interest quickly by greatly reducing overall human investigation time and effort.[5]

Mining writeprints from anonymous e-mails for forensic investigation is one of the paper for forensic data in which, three sets of experiments were performed by them which are (1) To assess stylometric attributes in terms of F-measure we applied clustering over nine different combinations of these attributes.(2) Other parameters will be constant for varying the number of authors. (3) In the third set of experiments the effects of number of messages per author had been checked by them. Three different clustering algorithms were used in all the three set of experiments, namely EM, k-means, and bisecting k-means were applied. {T1, T2, T3, T4, T1 þ T2, T1 þ T3, T2 þ T3, T1 þ T2 þ T3, T1 þ T2 þ T3 þ T4, } are Different feature combinations where T1, T2, T3 and T4 stand for lexical, syntactic, structural, and content-specific attributes in that order.

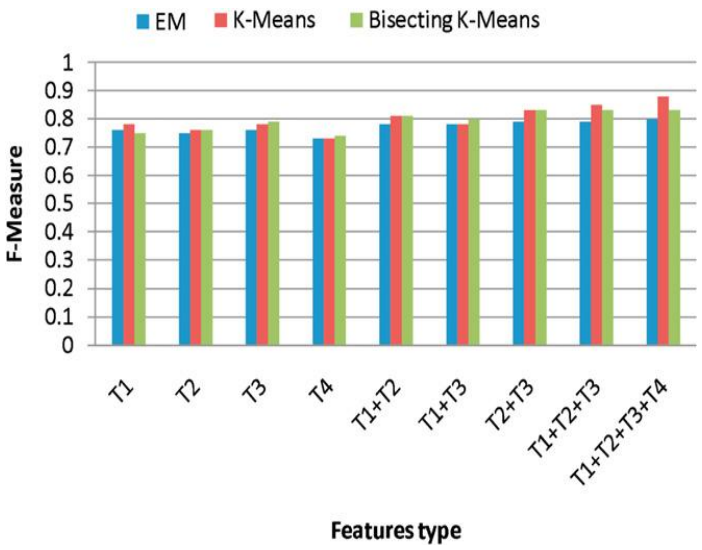

Fig. 1 E F-measure vs feature type and clustering algorithms

3. IMPLEMENTATION DETAILS:

3.1Architectural Diagram of Proposed System:

The proposed system shown in figure 2

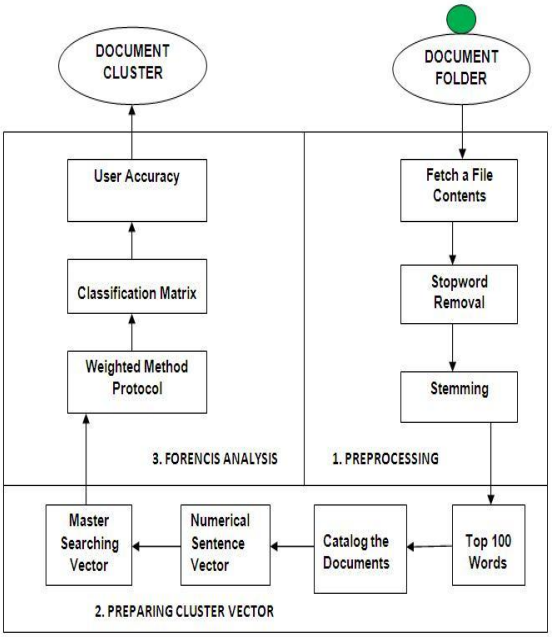

Fig. 2: Architectural diagram of proposed system

In our propose system basically there are three important steps which are as follows

1) Preprocessing
2) Preparing cluster vector
3) Forensic analysis

1) **Preprocessing-** In preprocessing step there are three steps such as a) fetch a file contents, b) stopword removal c) stemming. In all the above steps the basic purpose is to check the file contain and to remove the stop word like a, an ,the etc. and later on to do stemming on that file which will be removing ing and ed words from the given statement.

2) **Preparing Cluster Vector-** For preparing the cluster vector one will need to find top 100 words from the file on which preprocessing step is already done. Now from that document or rather way we can say file or data numerical sentences such as the sentence which has numerical word in it that means the sentence which contains date or any kind on number in it.

3) **Forensic Analysis-** This will be the last step of proposed method. From the diagram no 1 mention above one can say that for the forensic data analysis classification matrix need to be made with the help of weighted method protocol. At last one can find accuracy of his work.

4. RESULTS

4.1 Data Set

The data set for forensic analysis will be different number of file in different formant which has information on which data clustering is performed by applying dissimilar algorithm. For the clustering processes this paper makes use of multithreading technique. Later on that data set can be used for police investigation.

4.2 Result Set

The result set produced by this system will be number of clusters formed by applying algorithm on given information.

5. CONCLUSIONS AND FUTURE WORK

By doing the survey on computer forensic analysis it can be concluded that clustering on data is not an easy step. There is huge data to be cluster in compute forensic so to overcome this problem, this paper presented an approach that applies document clustering methods to forensic analysis of computers seized in police investigations. Again by using multithreading technique there will be document clustering for forensic data which will be useful for police investigations.

ACKNOWLEDGEMENTS

I express my sense of gratitude towards my project guide Prof. PREETI SHARMA for her valuable guidance at every step of study of this project, also her contribution for the solution of every problem at each stage.

I am also thankful to PROF. S. B. SARKAR, PG Coordinator for the motivation and Inspiration that triggered me for this thesis work.

REFERENCES

[1]. L. F. Nassif and E. R. Hruschka, "Document clustering for forensic computing: An approach for improving computer inspection," in Proc. Tenth Int. Conf. Machine Learning and Applications (ICMLA), 2011, vol. 1, pp. 265–268, IEEE Press.
[2]. L. Liu, J. Kang, J. Yu, and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering," in Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Engineering, 2005, pp. 597–601.
[3]. K. Stoffel, P. Cotofrei, and D. Han, "Fuzzy methods for forensic data analysis," in Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition, 2010, pp. 23–28.
[4]. R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis framework," Digital Investigation, Elsevier, vol. 5, no. 3–4, pp. 124–137, 2009.

[5]. B. K. L. Fei, J. H. P. Eloff, H. S. Venter, andM. S. Oliver, "Exploring forensic data with self-organizing maps," in Proc. IFIP Int. Conf. Digital Forensics, 2005, pp. 113–123.
[6]. F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," Digital Investigation, Elsevier, vol. 7, no. 1–2, pp. 56–64, 2010.
[7]. S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, "Text clustering for digital forensics analysis," Computat. Intell. Security Inf. Syst., vol. 63, pp. 29–36, 2009.